DOCUMENT RESUME

ED 388 712 TM 024 159

AUTHOR Wang, Xiang-bo; And Others
TITLE On the Viability of Some Untestable Assumptions in Equating Exams That Allow Examinee Choice. Program Statistics Research Technical Report No. 93-31.
INSTITUTION Educational Testing Service, Princeton, N.J.
SPONS AGENCY Graduate Record Examinations Board, Princeton, N.J.
REPORT NO ETS-RR-93-21
PUB DATE Mar 93
NOTE 19p.
PUB TYPE Reports - Research/Technical (143)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Adaptive Testing; Advanced Placement; *Difficulty Level; *Equated Scores; High Schools; *High School Students; History; *Multiple Choice Tests; Responses; Test Bias; *Test Format
IDENTIFIERS Advanced Placement Examinations (CEEB)

ABSTRACT

An increasingly popular test format allows examinees to choose the items they will answer from among a larger set. When examinee choice is allowed fairness requires that the different test forms thus formed be equated for their possible differential difficulty. For this equating to be possible it is necessary to know how well examinees would have answered the items that they did not choose. In this paper, results are reported for an experiment in which 213 high school students who took the Advanced Placement Chemistry examination were asked to choose among several multiple choice items but were then required to answer all of them. It is concluded that allowing choice while having fair tests is only possible when it is unnecessary. (Contains 3 tables, 5 figures, and 14 references.) (Author/SLD)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED 388 712

RR-93-21

On The Viability of Some Untestable Assumptions in Equating Exams That Allow Examinee Choice

Xiang-bo Wang
Educational Testing Service

Howard Wainer
Educational Testing Service

David Thissen
University of North Carolina

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
This document has been reproduced as received from the person or organization originating it.
Minor changes have been made to improve reproduction quality.
Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY
H. I. Braun
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)"

PROGRAM STATISTICS RESEARCH

TECHNICAL REPORT NO. 93-31

Educational Testing Service
Princeton, New Jersey 08541

TM024169

BEST COPY AVAILABLE

The Program Statistics Research Technical Report Series is designed to make the working papers of the Research Statistics Group at Educational Testing Service generally available. The series consists of reports by the members of the Research Statistics Group as well as their external and visiting statistical consultants.

Reproduction of any portion of a Program Statistics Research Technical Report requires the written consent of the author(s).

On the Viability of Some Untestable Assumptions in Equating Exams That Allow Examinee Choice

Xiang-Bo Wang
Educational Testing Service

Howard Wainer
Educational Testing Service

David Thissen
University of North Carolina

Program Statistics Research
Technical Report No. 93-31

Research Report No. 93-21

Educational Testing Service
Princeton, New Jersey 08541

March 1993

Copyright © 1993 by Educational Testing Service. All rights reserved.

On the viability of some untestable assumptions in equating exams that allow examinee choice§

Xiang-bo Wang & Howard Wainer
Educational Testing Service
&
David Thissen
University of North Carolina
at Chapel Hill

Abstract

An increasingly popular test format allows examinees to choose which items they will answer from among a larger set. When examinee choice is allowed fairness requires that the different test forms thus formed be equated for their possible differential difficulty. For this equating to be possible we need to know how well examinees would have answered the items that they did not choose. In this paper we report the results of an experiment in which examinees are asked to choose among several multiple choice items but are then required to answer all of them. We conclude that allowing choice while having fair tests is only possible when it is unnecessary.

§ This work was funded by a contract with the Graduate Record Examination Board and we are grateful for the opportunity to acknowledge this support. Our work has profitted from conversations with Paul Holland and Rick Morgan. Whatever clarity of exposition this paper possesses is due in no small degree to the careful review that it received from our colleague Rebecca Zwick, we are grateful for her help. The choice data we report was gathered by Xiang-bo Wang as part of his doctoral dissertation at the University of Hawai'i. This paper is the third in a series of collaborations on the same subject and the order of authorship has been permuted over this series.

On the viability of some untestable assumptions in equating exams that allow examinee choice

Introduction

An increasingly popular test format allows examinees to choose which items they will answer from among a larger set. This yields, in a very real sense, the possibility of many different examinee-created forms of the test. If comparisons are to be made among examinees who have taken different forms of a test, the canons of good practice require that those forms either be of equal difficulty or that the scoring procedure adjust statistically for the differences (equate the forms). In earlier work (Thissen Wainer, & Wang, 1992; Wainer, Wang, & Thissen, 1991), we report a methodology for accomplishing this adjustment as well as a case study illustrating it.

To accomplish such an equating requires assumptions about the distribution of scores for all of the items. Since we typically observe the scores for only those who opt to answer each choice item, the assumptions are aimed at the unobservable portion of this distribution: the scores that would have been observed on the choice items from the examinees who opted not to answer them. The viability of such assumptions are generally not testable. To our knowledge, there are no testing programs that equate choice items for their differential difficulty.

In this paper we provide:

1. a formal definition of the assumptions that we have made for equating the examinee-created test forms,

2. a description of data gathering methodology that allows them to be tested, and

3. an illustration, drawn from the College Board's Advanced Placement Exam in Chemistry, that indicates the extent to which these assumptions are upheld.

Do we really need to adjust?

Unequivocally "yes." There is enormous evidence supporting this conclusion (for expansion see Pomplun, Morgan & Nellikunnel, 1992 and Wainer & Thissen, 1992). We but skim the surface here.

In an unpublished technical memorandum, Fremer, Jackson & McPeek (1968) report unsettling results on a choice item from the 1968 AP Chemistry test. All examinees took a multiple choice section but had a choice between two largish problems; problem 4 and problem 5. Those who chose problem 4 were designated "Choice Group 1," those

who opted for problem 5 were called "Choice Group 2." Their performance on these two portions of the test are shown in Table 1.

Table 1

AP Chemistry 1968

(Fremer, Jackson & McPeek, 1968)

Choice Group	Mean Scores on Multiple Choice Section	Mean scores on Problem 4	Problem 5
1	11.7	8.2	
2	11.2		2.7

The observed difference on the multiple choice (MC) section is small. The difference on the choice items is large. While there are many possible explanations for these differences, the most plausible one is that the two choice problems are not equally difficult[1]. Fairness suggests that they ought to be equated.

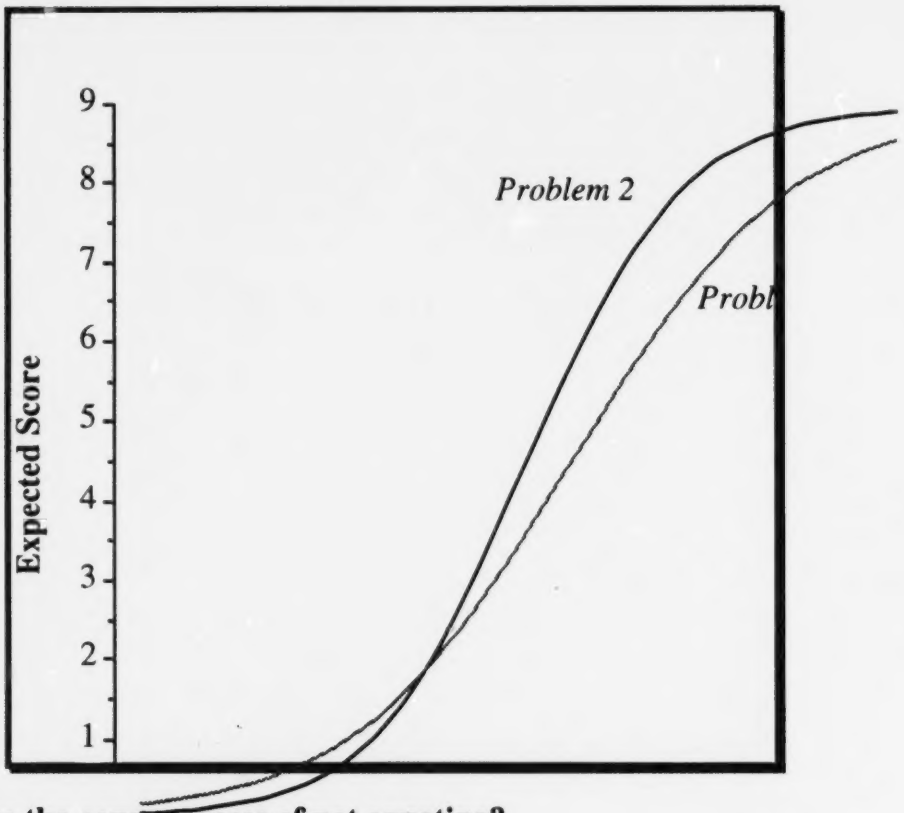
In the 1989 AP Chemistry test we showed (Wainer, Wang, & Thissen, 1991) that one pair of choice problems were not equally difficult. This is made apparent when one examines the expected score curves that characterize each item (see Figure 1). These curves are generated by calculating the expected score for an item under the polytomous IRT model that was fit to the data. Examinees who chose Problem 2 had as much as a one and a half point advantage over examinees of equal proficiency[2] who happened to choose problem 3.

---

[1]Two other explanations are: (i) multidimensionality, the problem is measuring something somewhat independent of the MC section, or (ii) the difference in the MC scores is , in fact, much larger than it appears. In this particular instance neither of these is true (see Thissen, Wainer & Wang, 1992)

[2]"Proficiency" is the term we use to characterize the IRT person parameter θ that is estimated when the entire test was fit by a single complex model. The details of this procedure are too lengthy to be rep[eated here but are described fully in Wainer, Wang & Thissen (1991).

Figure 1



What are the consequences of not equating?

Obviously, if we do not equate, those examinees who choose the more difficult items are adversely affected. Who are these unfortunates? In AP Chemistry and AP American History they tend to be women. In Table 2, below, are shown the results of examinee choice on section D of the 1989 AP Chemistry test. There are five problems on this section; labeled 5, 6, 7, 8, and 9. Examinees must choose three of these five problems. In Table 2 we see women tend to select the more difficult sets of problems. More women then men chose item sets that included the difficult problem 6, whereas men tended to chose item sets that include the easiest item (9) more often than women.[3]

[3]Interpretation of this table is helped by noting that about one-third of the approximately 18 thousand examinees who took this form of the test were women.

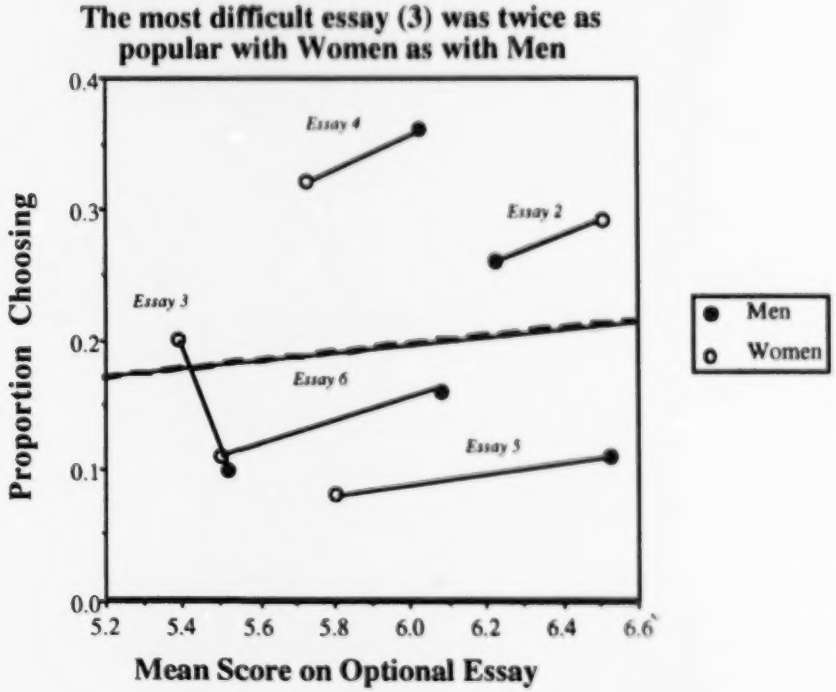
Table 2

Choice on Section D	Proportion Choosing *Males*	*Females*	
567	0.13	0.16	Hardest
568	0.25	0.35	
578	0.28	0.24	
789	0.06	0.03	
589	0.10	0.07	Easiest

Item 6 is the hardest
Items 8 & 9 are the easiest

On one portion of the 1989 AP American History exam, examinees are given five essay topics and asked to choose one to write on. In Figure 2 we see that the most difficult topic (essay 3) seems to attract a disproportionate number of women.

Figure 2



The practice of not equating for the differences in difficulty of choice items has the consequence, in these instances, of lowering women's scores relative to men. There is some evidence of a similar effect among various ethnic groups as well.

Assumptions for equating choice tests

The key assumption that we use for equating choice items is that nonresponse to the choice items is ignorable (Little & Rubin, 1987) conditioned on estimated proficiency. What we mean by this is that the probability of an examinee choosing any particular item is conditionally independent of her likelihood of getting that item correct, conditional on θ.

Stated more precisely, suppose

y_i is the score on test item Y_i, and

R_i is a choice function that takes the value 1 if Y_i is chosen and 0 if not.

In a choice situation we can observe the distribution of scores, $f_1(y)$, for those who opted to take an item. This can be denoted

$$f_1(y) = P(Y=y \mid R=1).$$

What we do not know, but which is crucial if we are to be able to equate the different choice items, is the distribution of scores, $f_0(y)$, for those who did not take the item. This is denoted

$$f_0(y) = P(Y=y \mid R=0).$$

To be able to do a proper equating we need to know the distribution of scores in the unselected population, $g(y) = P(Y=y)$.

Note that we can represent $g(y)$ as a mixture of two distributions

$$g(y) = f_1(y) \times P(R=1) + f_0(y) \times P(R=0).$$

The only unknown piece of this equation is $f_0(y)$, the distribution of scores among those individuals who chose not to answer it. Unless one engages in a special data gathering effort in which those examinees who did not answer Y_i, are forced to, $f_0(y)$ is not only unknown, it is unknowable. Thus the conundrum is that we must equate to insure fairness, but we cannot equate without knowing $f_0(y)$. What can we do?

One approach, mixture modeling (Glynn, Laird & Rubin, 1986) is to hypothesize a structure for $f_0(y)$ and proceed. In the past (Wainer, Wang & Thissen, 1991) we assumed that the function $f_0(y)$ is the same as $f_1(y)$.

In formal terms, this assumption is that

$$f_1(y) = P(Y=y \mid R=1, \theta) = P(Y=y \mid R=0, \theta) = f_0(y) = P(Y=y \mid \theta) \qquad (1)$$

Or in words, we assume that the traceline for the item is the same for those who didn't choose the item as it was for those that did. If we could gather the appropriate data (forcing responses from those who initially opted not to answer it) this hypothesis can easily be tested using standard DIF technology.

This sort of assumption of conditional independence has a surface similarity to the conditional independence assumption that underlies all of IRT, but is in fact, quite different. There is little cohesive evidence available that illuminates the relationship between examinees' preference for a particular item and their eventual scores. What little evidence there is, is indecisive. Thus, in the absence of contrary data and because this assumption does allow us to employ the existing machinery of IRT to equate, we use it.

Is this assumption empirically plausible?

As we pointed out, $f_0(y) = P(Y=y \mid R=0)$ is usually not observable and so the assumption that $f_0(y) = f_1(y)$ is not testable. However we engaged in a special data gathering effort that allows us to get a glimpse of the viability of this assumption.

Two hundred and thirteen (213) Advanced Placement Chemistry students in various high schools in Hawai'i took, as part of their final course exam, a special version of the 1989 AP Chemistry exam. Embedded within the 20 items of a multiple choice section were three pairs of choice questions. Prior to answering each pair of choice items each examinee was asked, "*If you were allowed to answer only one question, which would you choose?*" After responding, they were required to answer both items; the one they did not choose as well as the one they chose. This special design makes it possible to see what students' hypothetical scores would have been on the unchosen items on a choice test, and thus to test the normally untestable hypotheses that are required to equate tests formed through the exercise of examinee choice.

What did we find?

We designated the choice items embedded within the 20 multiple choice items of this test as item pairs (11, 12), (13, 14), and (15, 16). When each student reached these pairs they were instructed to indicate their preference before answering both of them. Shown in Table 2 are the empirical item difficulties[4] for these items and the frequency with which they were chosen.

---

[4]These difficulties are the 3-PL parameter *b* obtained from the 1989 operational administration of these items described in the 1990 College Board Report *The 1989 Advanced Placement Examination in Chemistry and its grading.* .

Table 2

The choice frequencies and difficulties
of the choice items

Item Number	Number Choosing	Difficulty *(b)*
11	180	-2.49
12	45	1.62
13	139	-0.04
14	83	0.00
15	78	2.09
16	143	2.12

Items 11 and 12 are very different in difficulty, with item 11 by far the easier for examinees in the operational sample. Note however that 45 students (20%) preferred to take that item. Perhaps, for them, this item was easier. Items 13 and 14 were both of middle difficulty and items 15 and 16 were both difficult.

Shown in Table 3 are the proportion of students who answered each of these items correctly, shown as a function of which item they chose. Note that those examinees who chose item 12 did far better on item 11. Thus if this choice had been part of an operational test these examinees' choice would have disadvantaged them. Adjusting for the difference in the difficulties of the two items would have ameliorated this somewhat. Note further that the students who chose item 11 did better on both items than those who chose item 12. This result is consonant with other research findings that suggest that better students are better able to judge item difficulty (Chi, Glaser & Rees, 1982; Chi, 1978).

The results from the other two item pairs are less striking. Examinees who chose item 13 appear somewhat more able than those who chose item 14, however they actually did marginally better on the item they originally rejected. Examinees who chose item 16 seem to show that there was little to choose between. But 17 ($.22 \times 78$) of those who chose item 15 seem to have made a mistake. These results indicate quite clearly that individuals do not always choose wisely, and hence allowing choice (at least in this situation) is not guaranteed to show the examinee in the best light.

Table 3

The proportion of examinees who answer each item correctly as a function of their choice

Item Answered	Item chosen 11	12
11	0.84	0.69
12	0.23	0.11

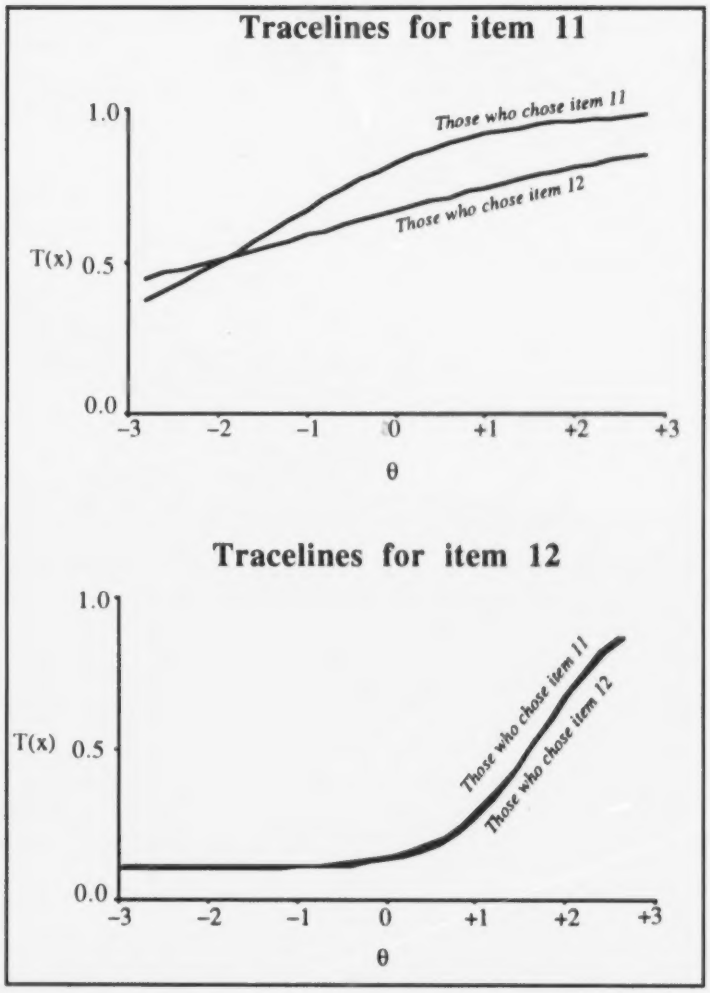
Item Answered	Item chosen 13	14
13	0.66	0.49
14	0.69	0.49

Item Answered	Item chosen 15	16
15	0.13	0.19
16	0.22	0.19

As we have observed in the 11-12 choice, unfairness was introduced into the test by examinee choice. This unfairness can be ameliorated if we adjust examinee scores for the differential difficulty of the items chosen. In order for us to make this adjustment we need to assume that $f_1(y) = P(Y=y \mid R=1, \theta) = P(Y=y \mid R=0, \theta) = f_0(y) = P(Y=y \mid \theta)$; that, for example, the traceline for item 11 is the same for those who chose item 11 as it would have been for those who chose item 12. To test these hypotheses we utilized IRT-based DIF technology (Thissen, Steinberg & Wainer, 1988, 1993; Wainer, Sireci, & Thissen, 1991)

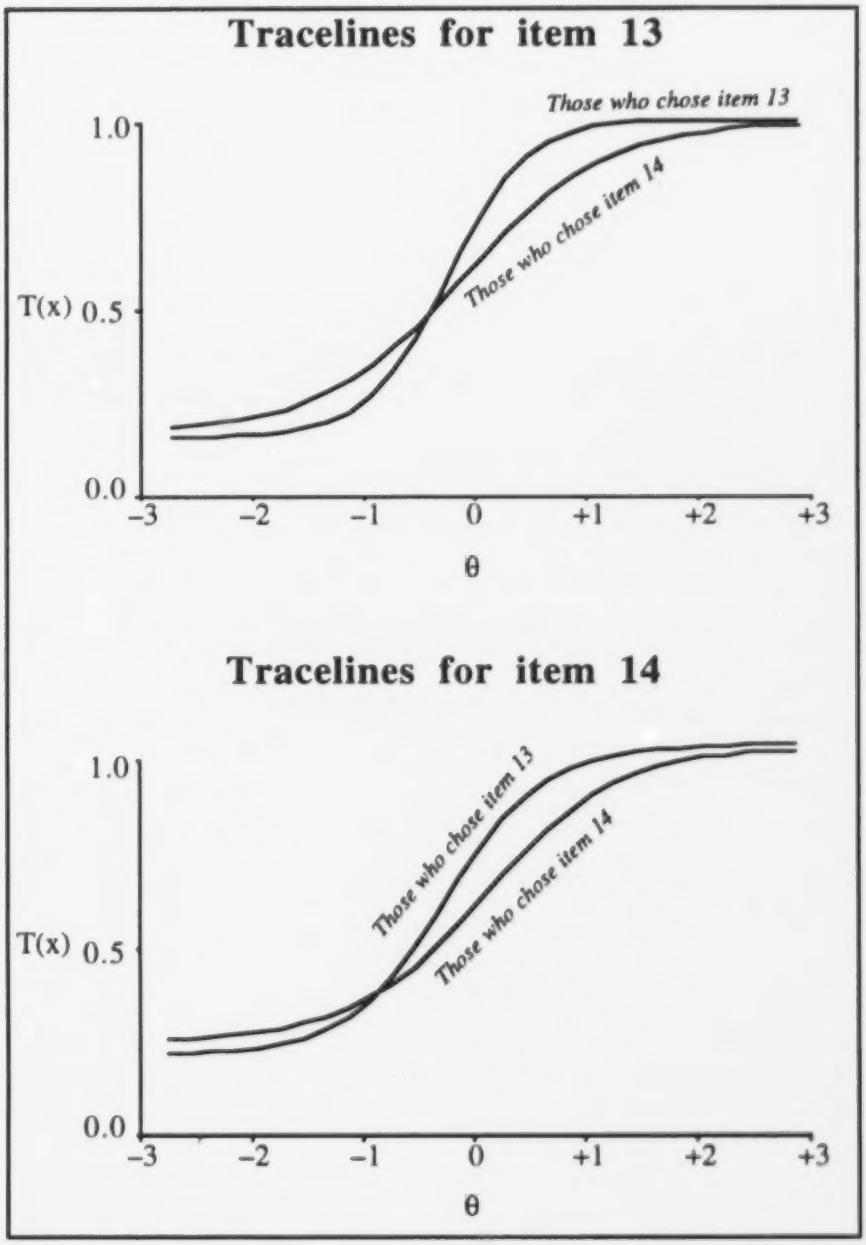
Figure 3 shows the estimated tracelines for choice items 11 and 12 for those examinees that chose each item as well as for those that did not. The apparent difference between the two tracelines for item 11 shows indications of being marginally significant ($\chi^2_{(2)} = 4$), whereas there is no difference between the two tracelines for item 12. Operationally this means the ordinarily untestable assumption that we used to equate choice forms may be untrue for item 11. Note that the differences observed in the traceline for item 11 suggest that item 11 is easier for those that chose it than for those examinees who did not.

Figure 3



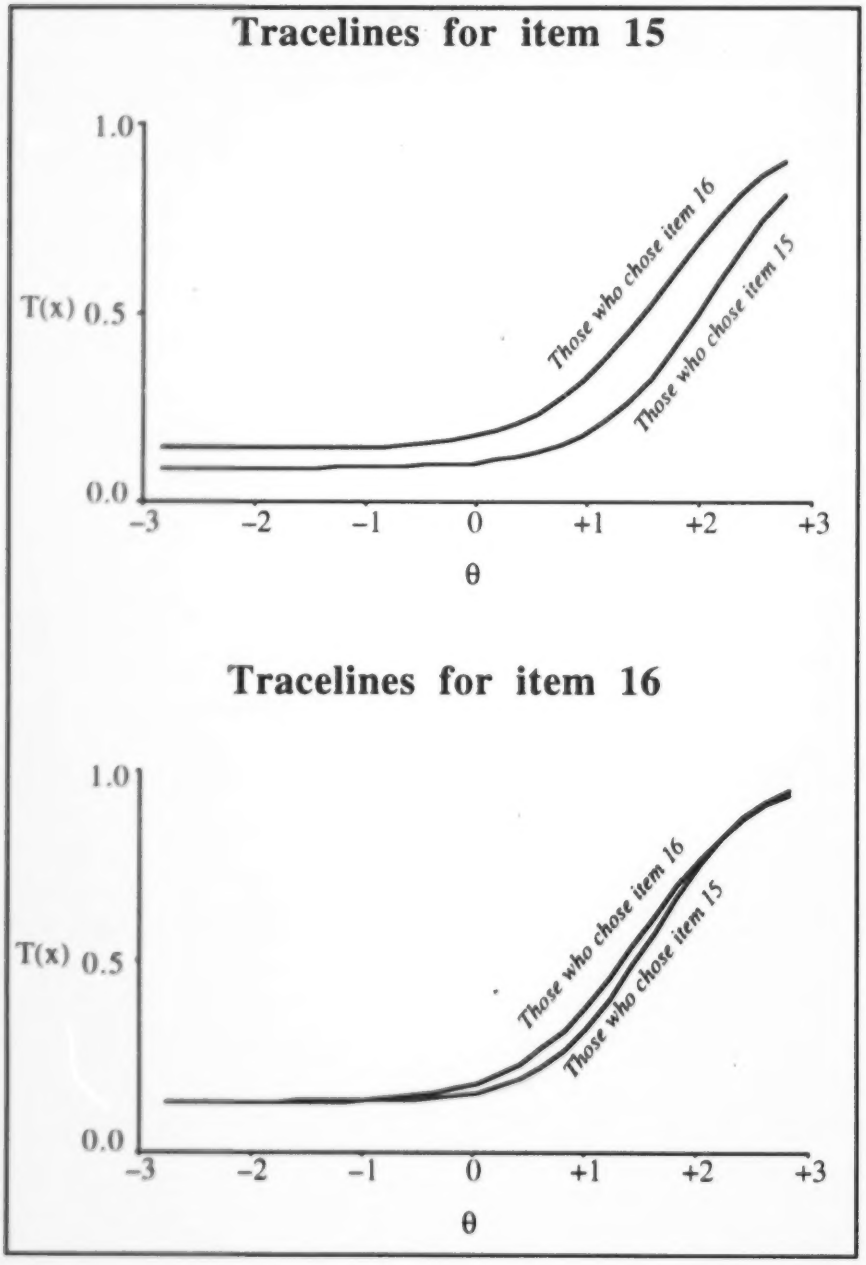
Looking at the tracelines for items 13 and 14, shown in figure 4 we see that they are quite similar, although not as similar as item 12's. These observed differences are not statistically significant.

Figure 4



Last are the tracelines for items 15 and 16 shown in figure 5. The observed difference between the two tracelines for item 15 is statistically significant, but in this instance it is easier for those who did not choose it. There is no difference between the tracelines for item 16.

Figure 5



Thus in an attempt to test the viability of the ignorability assumption that we need to equate choice items we have discovered that sometimes it works and sometimes it doesn't. Of the six items examined we found that four of the items performed identically among those who chose it as it did among those who did not. In the two instances in

which a difference was observed one time the group that chose the item did better on it than the group that didn't (item 11) and the other time the reverse was true (item 15).

Conclusions

The results are preliminary but discouraging. If examinees are allowed to choose the items they wish to answer from among a set of candidate items they are, in a very real sense, creating different test forms. These test forms are unlikely to be identical in difficulty. When different test forms are administered, it is considered important to equate those forms. Thus it seems sensible for us to want to equate forms that are generated by examinee choice. In order to accomplish this equating, the most plausible assumption is that the choice items have the same tracelines among those who chose it as they would have had among those who chose a different item. This assumption is closely allied to assuming ignorable nonresponse conditional on proficiency, which we found useful in earlier work (Wainer, Wang, & Thissen, 1991). In this small study we found that sometimes this assumption is viable and *sometimes it is not.* When it isn't the differences in the estimated tracelines for the choice items do not go in any predictable direction. We do not know the extent to which these results generalize to the sorts of large items that reflect what are the actual choice items in AP Chemistry. To the extent that these results do generalize puts a limit on the accuracy of the equating we did (Wainer, Wang, & Thissen, 1991).

What does this finding mean for practical testing with choice? There are two ways to interpret our results.

a. One way is to say that the differences we observed were smallish and that with proper pretesting it should be easy to find items whose performance is such that we can allow choice and still be able to equate for the differential difficulty of the choice items. That is we can allow choice if we are careful to only use items for which choice is irrelevant.

b. A second interpretation is one that recognizes the small sample size that we have used (only 45 examinees chose item 12)[5] and infers that with larger sample sizes the differences observed could have been much larger (because the data would overwhelm the priors in the estimation and hence yield posterior estimates considerably further away from one another) and would certainly have increased statistical significance. If this speculation turns out to be true it would mean that accurate equating of choice items could not take place and consequently we could not, in good conscience allow choice.

---

[5]How one estimates a stable ICC with only 45 examinees is something of a trick, the details of which are off the track in this paper. Obviously, it requires carefully chosen tight priors on the parameters within the estimation methodology. With such tight priors it takes real differences in the likelihood to pull the posterior estimates as far apart as we have found.

Putting these two together yields the discouraging conclusion that if we wish to have fair tests, choice is either unnecessary or impossible.

References

Author (1990). *The 1989 Advanced Placement Examination in Chemistry and its grading.* New York: The College Board.

Chi, M. T. H. (1978). Knowledge structures and memory development. In R. S. Siegler (Ed.), *Children's thinking: What develops?* Hillsdale, NJ: Erlbaum.

Chi, M. T. H., Glaser, R., & Rees, E. (1982). Expertise in problem solving. In R. J. Sternberg (Ed.), *Advances in the Psychology of Human Intelligence.* Hillsdale, NJ: Erlbaum.

Fremer, J., Jackson, R., & McPeek, M. (1968). *Review of the psychometric characteristics of the Advanced Placement Tests in Chemistry, American History, and French.* Internal Memorandum. Princeton, NJ: Educational Testing Service.

Glynn, R. J., Laird, N. M., and Rubin, D. B. (1986) Selection Modeling versus mixture modeling with nonignorable nonresponse. In H. Wainer (ed.) *Drawing Inferences from Self-Selected Samples.* New York: Springer-Verlag, pps. 115- 142.

Little, R. J. A., & Rubin, D. B. (1987). *Statistical analysis with missing data.* New York: Wiley.

Pomplun, M., Morgan, R., & Nellikunnel, A. (1992). *Choice in Advanced Placement Tests,* Unpublished Statistical Report (SR-92-51). Princeton, NJ: Educational Testing Service.

Thissen, D., Steinberg, L. & Wainer, H. (1988). Use of item response theory in the study of group differences in trace lines. In H. Wainer & H. Braun (Eds.), *Test Validity.* Hillsdale, NJ: Erlbaum, pp. 147-169.

Thissen, D., Steinberg, L. & Wainer, H. (1993). Detection of *differential item functioning* using the parameters of item response models. In P.W. Holland & H. Wainer (Eds.), *Differential item functioning.* Hillsdale, NJ: Lawrence Erlbaum Associates, Chapter 4, pps. 67-113.

Thissen, D., Wainer, H. & Wang, X-B. (1992). *How unidimensional are tests comprising both Multiple-choice and Free-Response Items? An analysis of two tests.* ETS Technical Report (92-xx). Princeton, NJ.: Educational Testing Service.

Wainer, H. & Thissen, D. (1992). *On examinee choice in educational testing.* ETS Technical Report (92-xx). Princeton, NJ.: Educational Testing Service.

Wainer, H. & Thissen, D. (1992). *Choosing: A Test.* ETS Technical Report (92-xx). Princeton, NJ.: Educational Testing Service.

Wainer, H., Sireci, S.G. & Thissen, D. (1991). Differential testlet functioning: Definitions and detection. *Journal of Educational Measurement, 28*, 197-219.

Wainer, H., Wang, X. B., & Thissen, D. (1991). *How well can we equate test forms that are constructed by examinees?* Technical Report (TR-91-15). Princeton, NJ: Educational Testing Service.